Introduction to the OLCF and OLCF-3

OLCF

Oak Ridge Leadership Computing Facility

Presented by:

Bronson Messer

Acting Director of Science

Oak Ridge Leadership Computing Facility

ORNL has a long history in High Performance Computing

ORNL has had 20 systems

on the TOP500® SUPERCOMPUTER SITES lists

1954
ORACLE

1969
IBM 360/9

1985
Cray X-MP

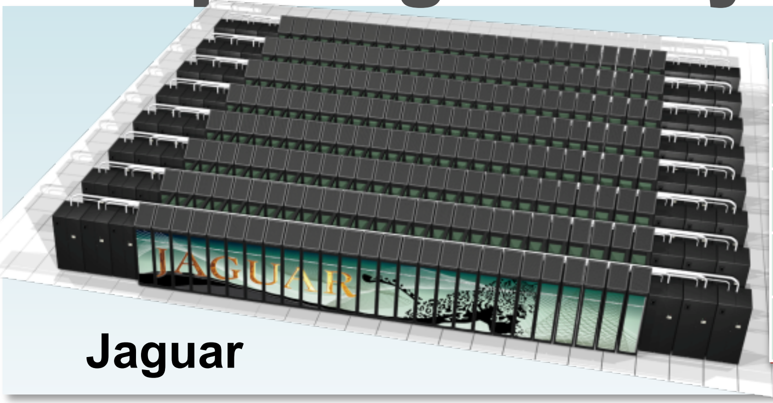
1992-1995
Intel Paragons

1996-2002
IBM Power 2/3/4

2003-2005
Cray X1/X1E

2007
IBM Blue Gene/P

Today, we have the world's most powerful computing facility

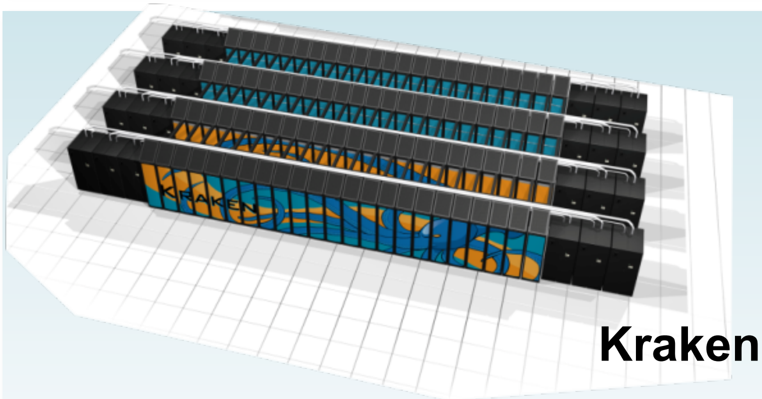


Jaguar

Peak performance	2.33 PF/s
Memory	300 TB
Disk bandwidth	> 240 GB/s
Square feet	5,000
Power	7 MW

TOP500 SUPERCOMPUTER SITES #2

Dept. of Energy's
most powerful computer

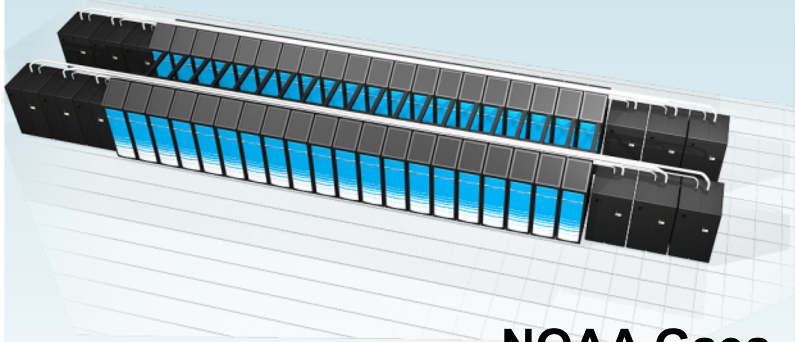


Kraken

Peak performance	1.03 PF/s
Memory	132 TB
Disk bandwidth	> 50 GB/s
Square feet	2,300
Power	3 MW

TOP500 SUPERCOMPUTER SITES #8

National Science
Foundation's most
powerful computer



NOAA Gaea

Peak Performance	1.1 PF/s
Memory	248 TB
Disk Bandwidth	104 GB/s
Square feet	1,600
Power	2.2 MW

TOP500 SUPERCOMPUTER SITES #32

National Oceanic and
Atmospheric Administration's
most powerful computer

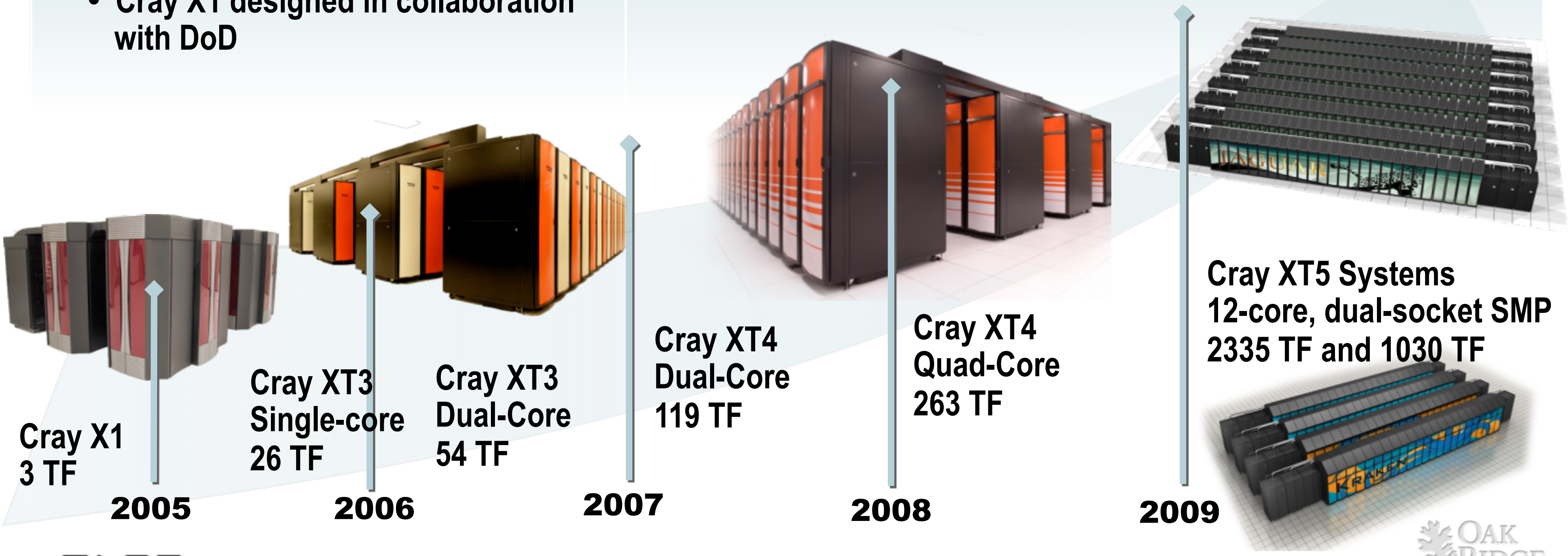
We have increased system performance by 1,000 times since 2004

Hardware scaled from single-core through dual-core to quad-core and dual-socket , 12-core SMP nodes

- NNSA and DoD have funded much of the basic system architecture research
  - Cray XT based on Sandia Red Storm
  - IBM BG designed with Livermore
  - Cray X1 designed in collaboration with DoD

Scaling applications and system software is the biggest challenge

- DOE SciDAC and NSF PetaApps programs are funding scalable application work, advancing many apps
- DOE-SC and NSF have funded much of the library and applied math as well as tools
- Computational Liaisons key to using deployed systems



Cray X1
3 TF
2005

Cray XT3
Single-core
26 TF
2006

Cray XT3
Dual-Core
54 TF

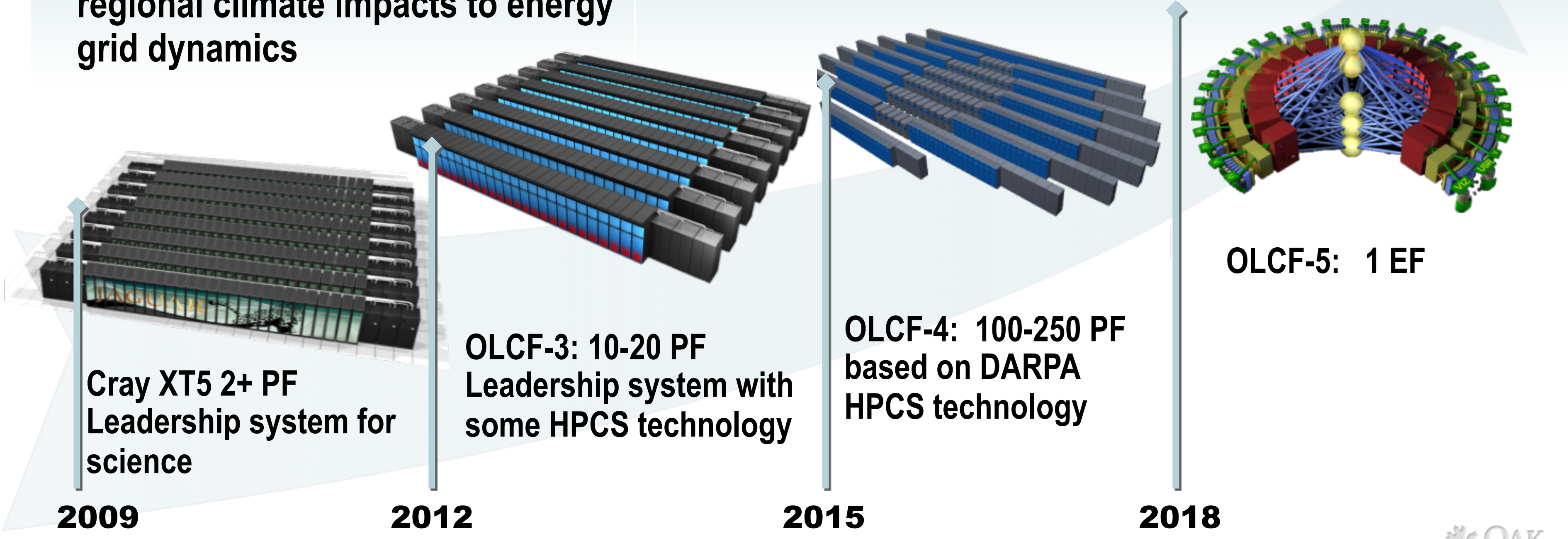
Cray XT4
Dual-Core
119 TF
2007

Cray XT4
Quad-Core
263 TF
2008

Cray XT5 Systems
12-core, dual-socket SMP
2335 TF and 1030 TF
2009

Our science requires that we advance computational capability 1000x over the next decade

Mission: Deploy and operate the computational resources required to tackle global challenges	Vision: Maximize scientific productivity and progress on the largest scale computational problems
• Deliver transforming discoveries in climate, materials, biology, energy technologies, etc. • Ability to investigate otherwise inaccessible systems, from regional climate impacts to energy grid dynamics	• Providing world-class computational resources and specialized services for the most computationally intensive problems • Providing stable hardware/software path of increasing scale to maximize productive applications development



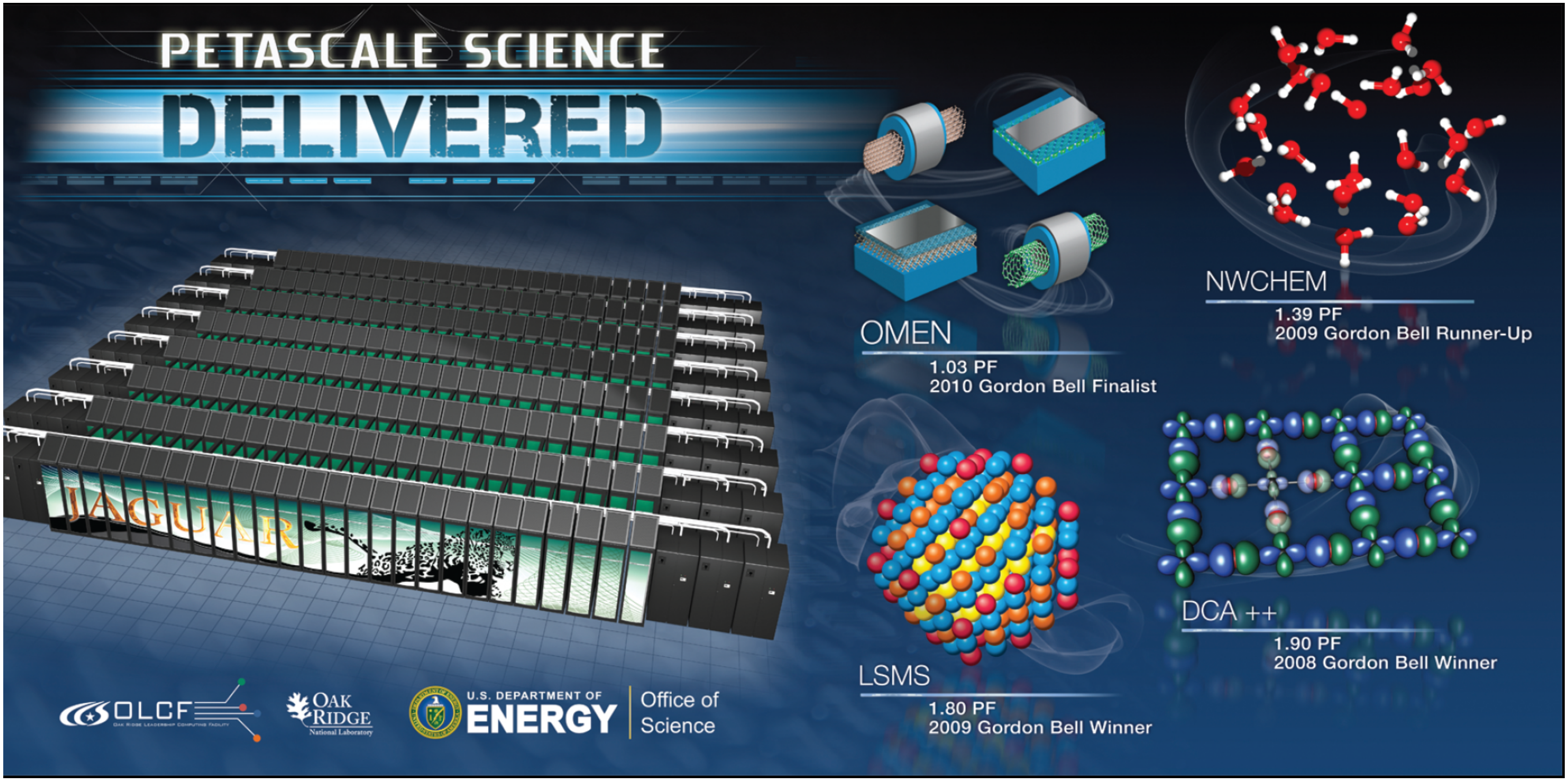
Cray XT5 2+ PF Leadership system for science

OLCF-3: 10-20 PF Leadership system with some HPCS technology

OLCF-4: 100-250 PF based on DARPA HPCS technology

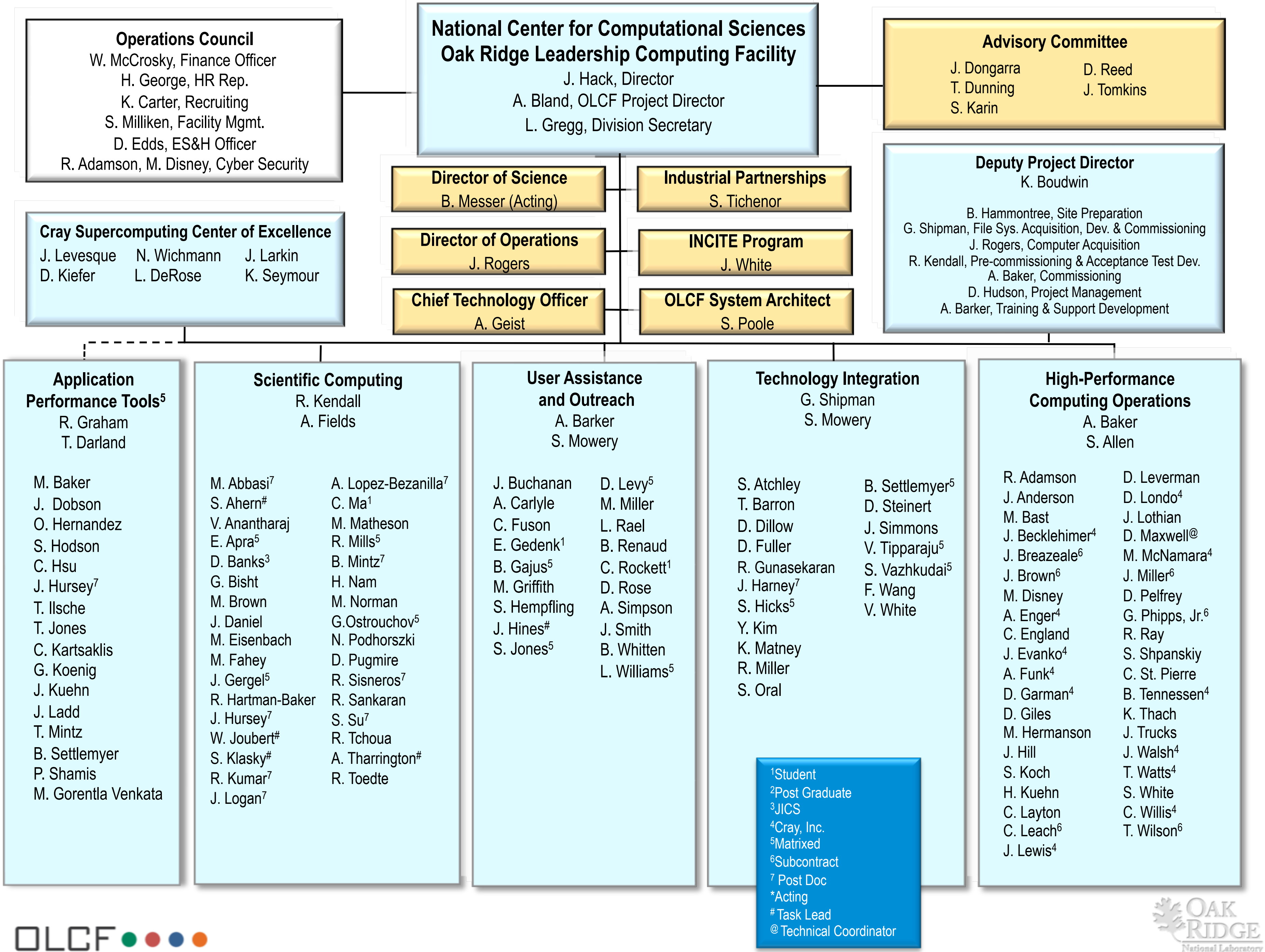
OLCF-5: 1 EF

We are delivering Petascale Science Today!



We have worked with science teams to scale codes to use Jaguar's 224,256 cores

Science Area	Code	Contact	Cores	Total Performance	Notes
Materials	DCA++	Schulthess	213,120	1.9 PF*	2008 Gordon Bell Winner
Materials	WL-LSMS	Eisenbach	223,232	1.8 PF	2009 Gordon Bell Winner
Chemistry	NWChem	Apra	224,196	1.4 PF	2009 Gordon Bell Finalist
Materials	DRC	Schulthess	186,624	1.3 PF	2010 Gordon Bell Honorable Mention
Nanoscience	OMEN	Klimeck	222,720	1.03 PF	
Biomedical	MoBo	Biros	196,608	780 TF	2010 Gordon Bell Winner
Seismology	SPECFEM3D	Carrington	149,784	165 TF	2008 Gordon Bell Finalist
Weather	WRF	Michalakes	150,000	50 TF	
Combustion	S3D	Chen	144,000	83 TF	
Fusion	GTC	PPPL	102,000	20 billion Particles / sec	
Materials	LS3DF	Wang	147,456	442 TF	2008 Gordon Bell Winner
Chemistry	MADNESS	Harrison	140,000	550+ TF	



National Center for Computational Sciences
Oak Ridge Leadership Computing Facility

J. Hack, Director
A. Bland, OLCF Project Director
L. Gregg, Division Secretary

Operations Council
W. McCrosky, Finance Officer
H. George, HR Rep.
K. Carter, Recruiting
S. Milliken, Facility Mgmt.
D. Edds, ES&H Officer
R. Adamson, M. Disney, Cyber Security

Advisory Committee
J. Dongarra
T. Dunning
S. Karin
D. Reed
J. Tomkins

Cray Supercomputing Center of Excellence
J. Levesque
D. Kiefer
N. Wichmann
L. DeRose
J. Larkin
K. Seymour

Director of Science
B. Messer (Acting)

Director of Operations
J. Rogers

Chief Technology Officer
A. Geist

Industrial Partnerships
S. Tichenor

INCITE Program
J. White

OLCF System Architect
S. Poole

Deputy Project Director
K. Boudwin

B. Hammontree, Site Preparation
G. Shipman, File Sys. Acquisition, Dev. & Commissioning
J. Rogers, Computer Acquisition
R. Kendall, Pre-commissioning & Acceptance Test Dev.
A. Baker, Commissioning
D. Hudson, Project Management
A. Barker, Training & Support Development

Application Performance Tools[5]
R. Graham
T. Darland

M. Baker
J. Dobson
O. Hernandez
S. Hodson
C. Hsu
J. Hursey[7]
T. Ilsche
T. Jones
C. Kartsaklis
G. Koenig
J. Kuehn
J. Ladd
T. Mintz
B. Settlemyer
P. Shamis
M. Gorentla Venkata

Scientific Computing
R. Kendall
A. Fields

M. Abbasi[7]
S. Ahern#
V. Anantharaj
E. Apra[5]
D. Banks[3]
G. Bisht
M. Brown
J. Daniel
M. Eisenbach
M. Fahey
J. Gergel[5]
R. Hartman-Baker
J. Hursey[7]
W. Joubert#
S. Klasky#
R. Kumar[7]
J. Logan[7]
A. Lopez-Bezanilla[7]
C. Ma[1]
M. Matheson
R. Mills[5]
B. Mintz[7]
H. Nam
M. Norman
G.Ostrouchov[5]
N. Podhorszki
D. Pugmire
R. Sisneros[7]
R. Sankaran
S. Su[7]
R. Tchoua
A. Tharrington#
R. Toedte

User Assistance and Outreach
A. Barker
S. Mowery

J. Buchanan
A. Carlyle
C. Fuson
E. Gedenk[1]
B. Gajus[5]
M. Griffith
S. Hempfling
J. Hines#
S. Jones[5]
D. Levy[5]
M. Miller
L. Rael
B. Renaud
C. Rockett[1]
D. Rose
A. Simpson
J. Smith
B. Whitten
L. Williams[5]

Technology Integration
G. Shipman
S. Mowery

S. Atchley
T. Barron
D. Dillow
D. Fuller
R. Gunasekaran
J. Harney[7]
S. Hicks[5]
Y. Kim
K. Matney
R. Miller
S. Oral
B. Settlemyer[5]
D. Steinert
J. Simmons
V. Tipparaju[5]
S. Vazhkudai[5]
F. Wang
V. White

High-Performance Computing Operations
A. Baker
S. Allen

R. Adamson
J. Anderson
M. Bast
J. Becklehimer[4]
J. Breazeale[6]
J. Brown[6]
M. Disney
A. Enger[4]
C. England
J. Evanko[4]
A. Funk[4]
D. Garman[4]
D. Giles
M. Hermanson
J. Hill
S. Koch
H. Kuehn
C. Layton
C. Leach[6]
J. Lewis[4]
D. Leverman
D. Londo[4]
J. Lothian
D. Maxwell@
M. McNamara[4]
J. Miller[6]
D. Pelfrey
G. Phipps, Jr.[6]
R. Ray
S. Shpanskiy
C. St. Pierre
B. Tennessen[4]
K. Thach
J. Trucks
J. Walsh[4]
T. Watts[4]
S. White
C. Willis[4]
T. Wilson[6]

[1]Student
[2]Post Graduate
[3]JICS
[4]Cray, Inc.
[5]Matrixed
[6]Subcontract
[7] Post Doc
*Acting
# Task Lead
@ Technical Coordinator

Evolution of Computer Architectures

1950s to 1960s: Architectural innovations for faster processors

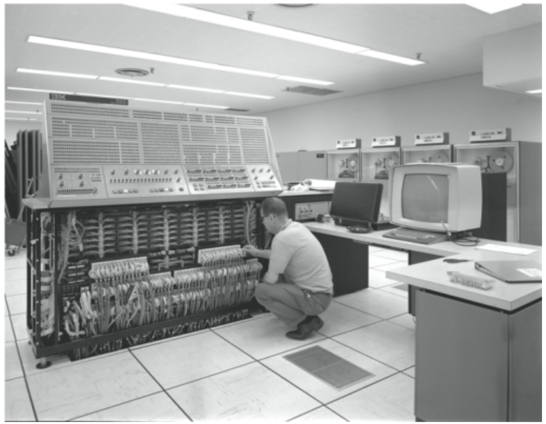
1970s: Vectors to increase performance

1980s: Parallelism begins

1990s: Era of massive parallelism and clock rate scaling

2000s: Parallelism continues, but clock rate scaling ends. Multicore processors

2010s: Multicore and accelerator based systems push parallelism from O(100K) to O(10B) way parallelism



IBM 360/9

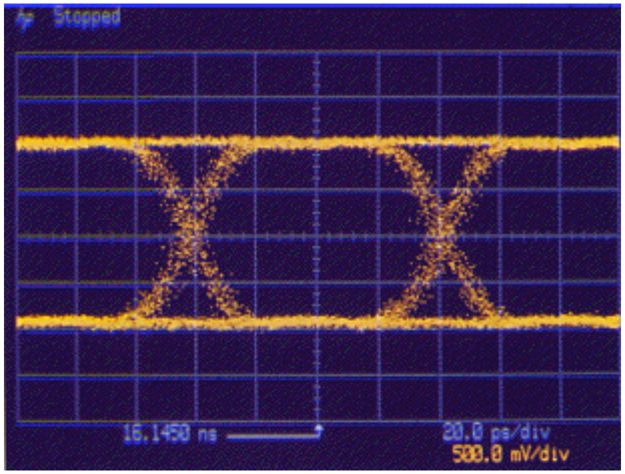
Cray X-MP

Intel Paragon

Why has clock rate scaling ended?

Power = Capacitance * Frequency * Voltage^2 + Leakage

- Traditionally, as Frequency increased, Voltage decreased, keeping the total power in a reasonable range
- But we have run into a wall on voltage
  - As the voltage gets smaller, the difference between a “one” and “zero” gets smaller. Lower voltages mean more errors.
  - While we like to think of electronics as digital devices, inside we use analog voltages to represent digital states.
- Capacitance increases with the complexity of the chip
- Total power dissipation is limited by cooling



Power to move data

$$\mathit{Energy\_to\_move\_data} = \mathit{bitrate} * \mathit{length}^2 \ / \ \mathit{cross\_section\_area\_of\_wire}$$

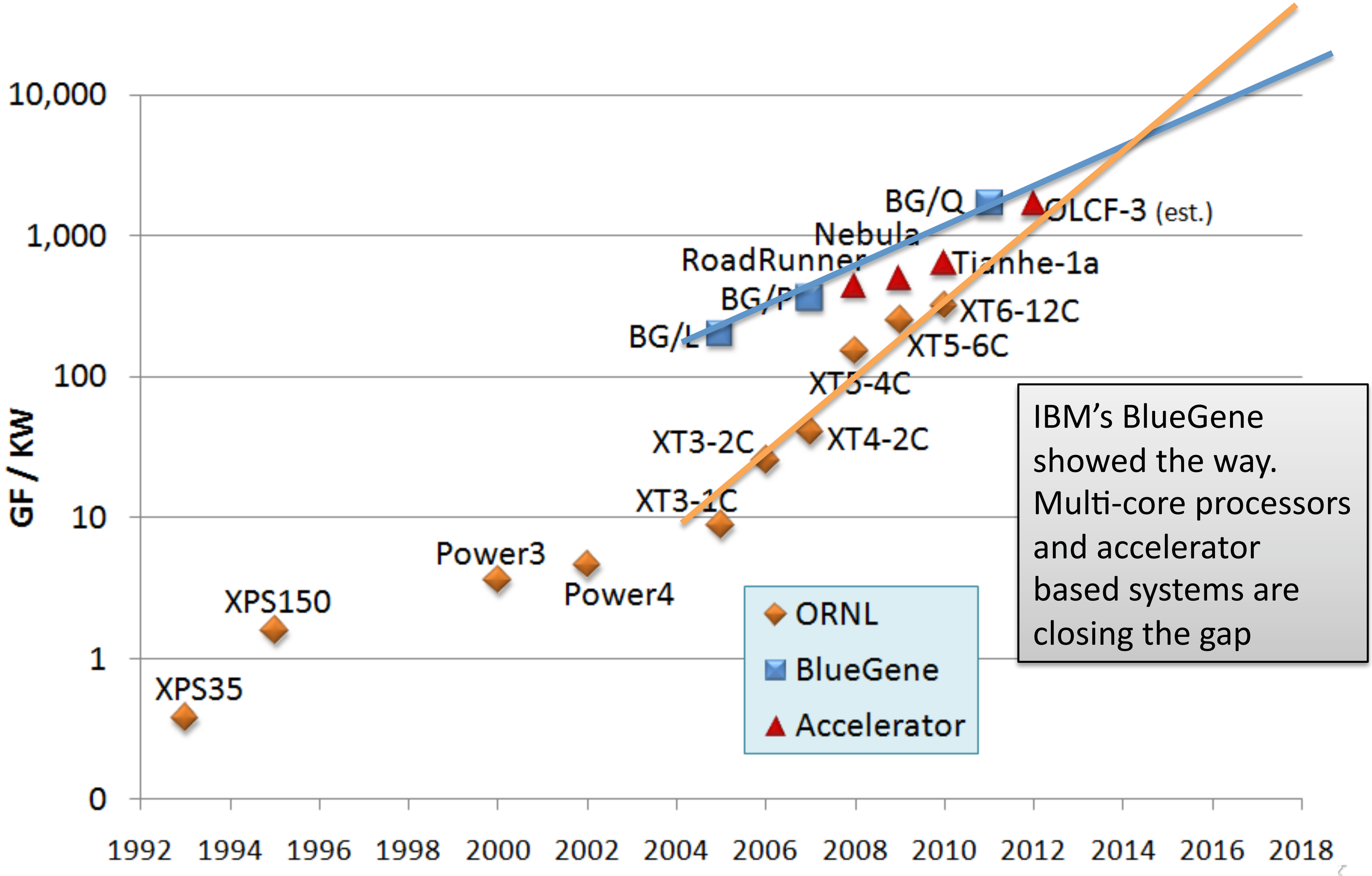
- The energy consumed increases proportionally to the bit-rate, so as we move to ultra-high-bandwidth links, the power requirements will become an increasing concern.
- The energy consumption is highly distance-dependent (the square of the length term), so bandwidth is likely to become increasingly localized as power becomes a more difficult problem.
- Improvements in chip lithography (making smaller wires) will not improve the energy efficiency or data carrying capacity of electrical wires.

D. A. B. Miller and H. M. Ozaktas, “Limit to the Bit-Rate Capacity of Electrical Interconnects from the Aspect Ratio of the System Architecture,” Journal of Parallel and Distributed Computing, vol. 41, pp. 42-52 (1997) article number PC961285.

Implications for future systems

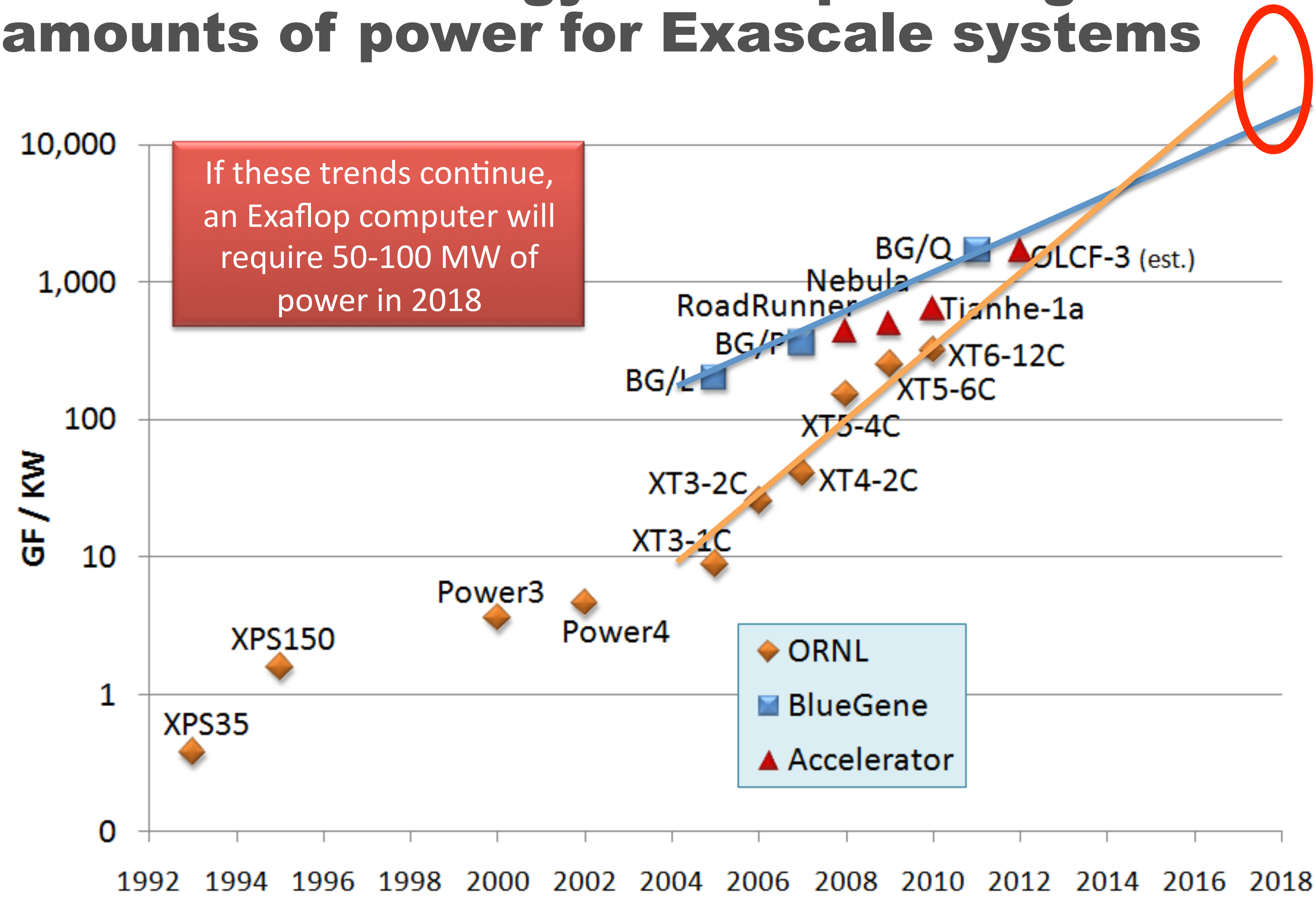
- Clock rates will stay largely the same as today, increasing the parallelism of systems to improve performance
- Energy cost of moving data is very large. We will have to explicitly manage data locality to limit power consumption

Trends in power efficiency



IBM's BlueGene showed the way. Multi-core processors and accelerator based systems are closing the gap

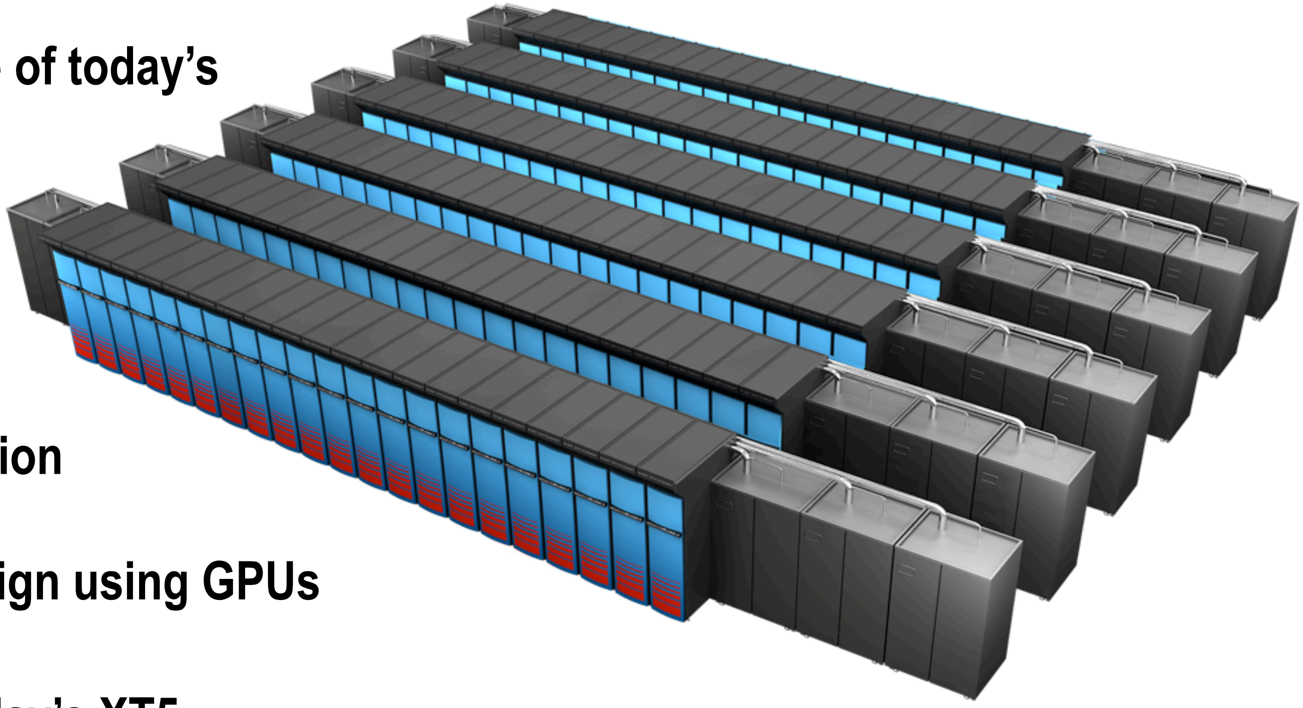
Current Technology will require huge amounts of power for Exascale systems



If these trends continue, an Exaflop computer will require 50-100 MW of power in 2018

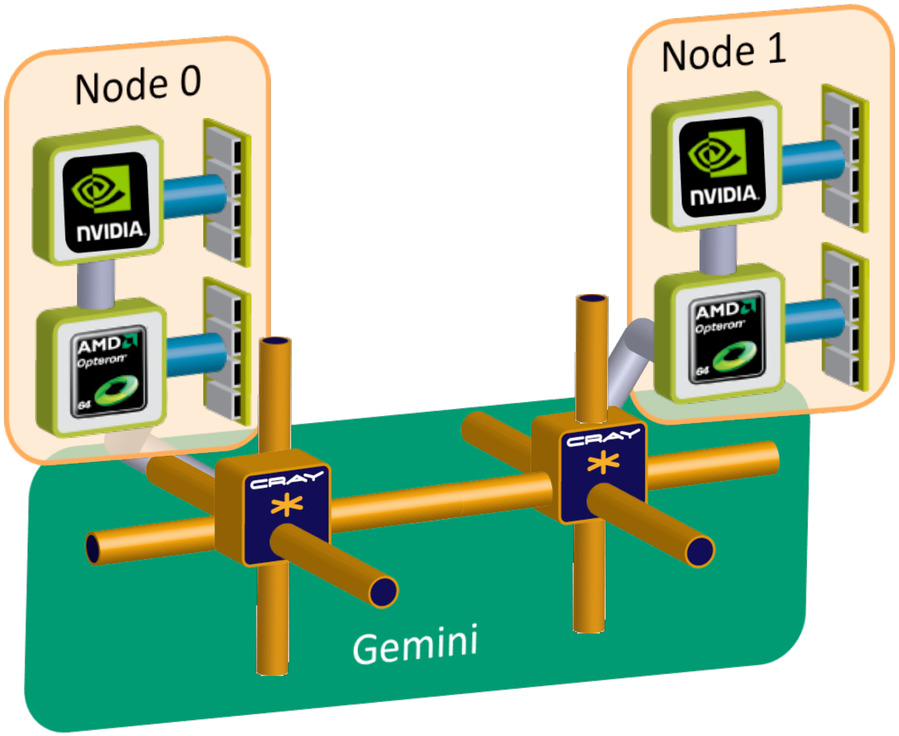
ORNL's "Titan" 20 PF System Goals

- Initial 1 PF delivery in 2011, final 20 PF system in 2012
- Designed for science from the ground up
- Similar number of cabinets, cabinet design, and cooling as Jaguar
- Operating system upgrade of today's Linux Operating System
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- New accelerated node design using GPUs
- 20 PF peak performance
  - 9x performance of today's XT5
- Larger memory
- 3x larger and 4x faster file system



OLCF-3 node description

- New node for "Cray XE" infrastructure
  - Gemini interconnect
  - AMD Socket G34 processor
- 1 AMD socket G34 processor and 1 NVIDIA GPU per node
- Interlagos uses AMD socket G34 and new "Bulldozer" core
  - DDR3-1600 memory
  - HyperTransport version 3
- NVIDIA "Kepler" accelerator
  - Successor to Fermi



	Jaguar's XT5 node	OLCF-3 node
Opteron sockets	2	1
Opteron memory (GB)	16	32
Interconnect	Seastar2	Gemini
Node peak GFLOPS	110	>1500

Why use an accelerator?

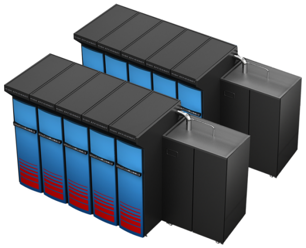
- Best way to get to a very powerful node
  - Titan nodes will be greater than 1.5 TeraFLOPS per node
- Power consumption per GF is much better than a conventional processor

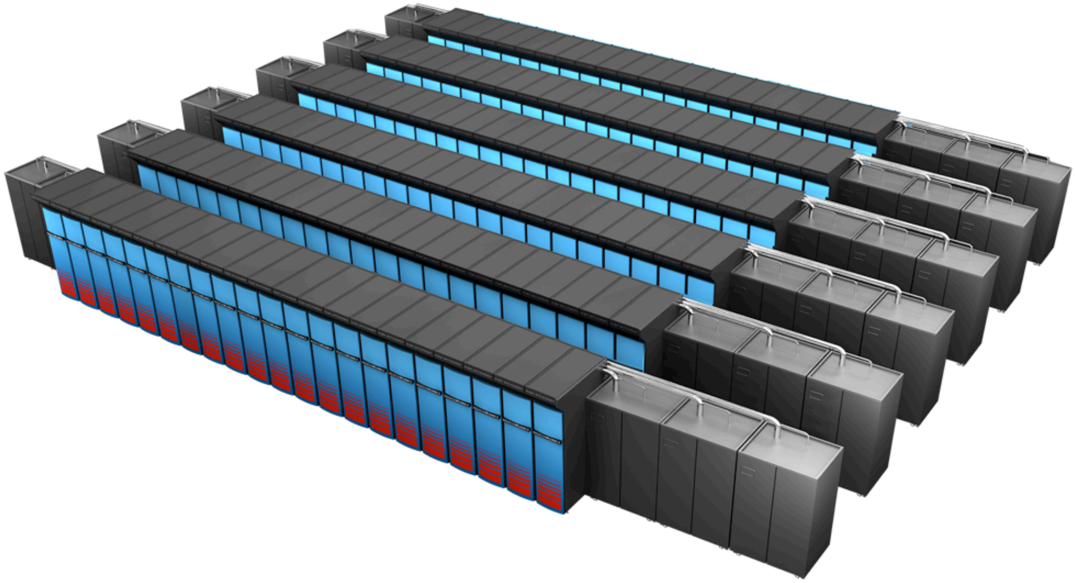
Processor type	GigaFLOPS / Watt
Cray XE6 (Magny-Cours)	1
Titan (Projected)	6.3

- Explicitly managed memory hierarchy
  - Programmer places the data in the appropriate memory and manages to save energy

OLCF-3 hardware plan maximizes science output

Initial Delivery System (IDS)	Final System	Scalable File System
• 2nd half of 2011 • 900 TF peak • 10 cabinets • 920 compute nodes • Prepare applications, mitigate system software risk	• 2nd half of 2012 • Incorporates upgraded IDS • 16–20 PF peak • 132 cabinets • 12,160 compute nodes	• Expansion of Spider • Adds 400–700 GB/s of bandwidth • Adds 10–30 PB of disk capacity





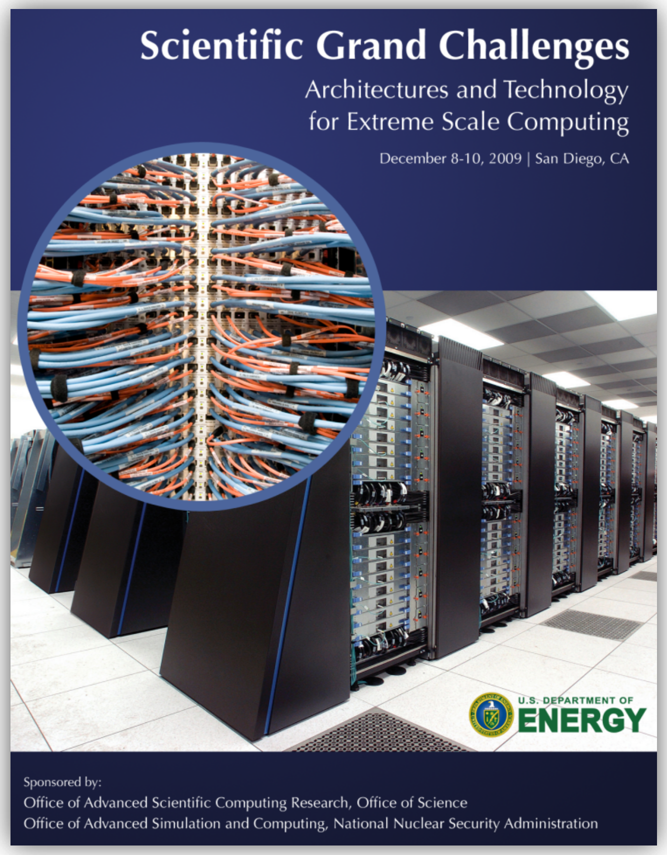
File System

- We will continue to use Lustre for our file system for Titan
- Plan to use Lustre version 2.x
  - Much more scalable metadata
  - Etc. etc. etc.
- Competitive procurement for the storage
  - Expect to get between 400 and 700 Gigabytes per second of bandwidth
  - Expect to add between 10 and 30 Petabytes of storage

Outline

- Drivers for HPC architectural change
- OLCF-3 “Titan” Overview
- **Programming model for highly parallel systems**

But is this enough as we look to exascale systems?



Scientific Grand Challenges

Architectures and Technology for Extreme Scale Computing

December 8-10, 2009 | San Diego, CA

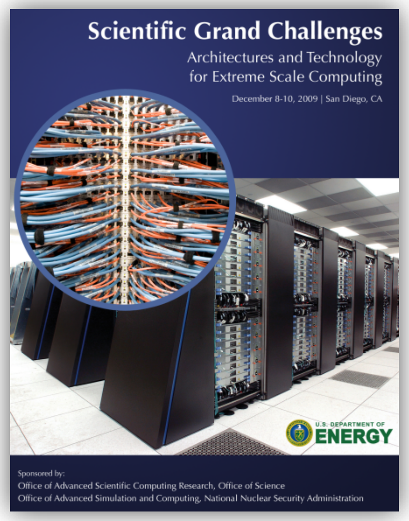
U.S. DEPARTMENT OF ENERGY

Sponsored by:
Office of Advanced Scientific Computing Research, Office of Science
Office of Advanced Simulation and Computing, National Nuclear Security Administration

- “Node architectures are expected to change dramatically in the next decade, becoming more hierarchical and heterogeneous.”
- “. . . computer companies are dramatically increasing on-chip parallelism to improve performance. The traditional doubling of clock speeds every 18 to 24 months is being replaced by a doubling of cores or other parallelism mechanisms.”
- “Systems will consist of one hundred thousand to one million nodes and perhaps as many as a billion cores.”

Architectures and Technology for Extreme Scale Computing, Workshop Report, 2009; http://www.er.doe.gov/ascr/ProgramDocuments/Docs/Arch-TechGrandChallengesReport.pdf

What does this say about the programming model?



Scientific Grand Challenges

Architectures and Technology for Extreme Scale Computing

December 8-10, 2009 | San Diego, CA

U.S. DEPARTMENT OF ENERGY

Sponsored by:
Office of Advanced Scientific Computing Research, Office of Science
Office of Advanced Simulation and Computing, National Nuclear Security Administration

- “The principal programming environment challenges will be on the exascale node: concurrency, hierarchy and heterogeneity.”
- “. . . more than a billion-way parallelism to fully utilize an exascale system”
- “Portability will be a significant concern . . . In order to improve productivity a programming model that abstracts some of the architectural details from software developers is highly desirable.”

Architectures and Technology for Extreme Scale Computing, Workshop Report, 2009; http://www.er.doe.gov/ascr/ProgramDocuments/Docs/Arch-TechGrandChallengesReport.pdf

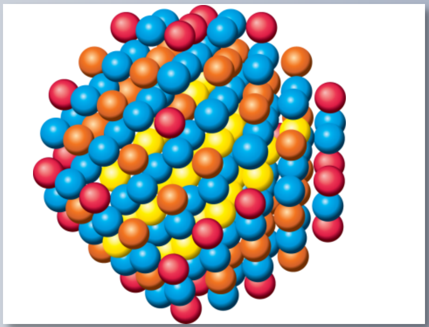
What should the programming model look like?

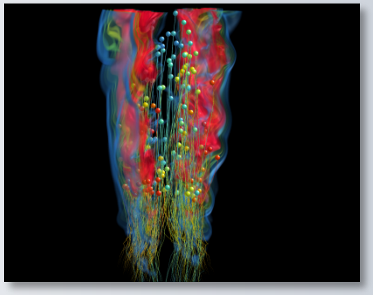
1. MPI or Global Address Space languages across nodes
2. Within the very powerful nodes, use OpenMP, or other threads package to exploit the large number of cores
3. In each thread, use directives to invoke vector, SIMD, or SSE style instructions in the processor or accelerator to maximize performance
4. Explicitly manage data movement to minimize power
5. Describe the parallelism in the high-level language in a portable way, then let the compiler and libraries generate the best code for the architecture

We are implementing this programming model on Titan, but this model works on current and future systems

We selected six early science applications to port to this architecture

WL-LSMS
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

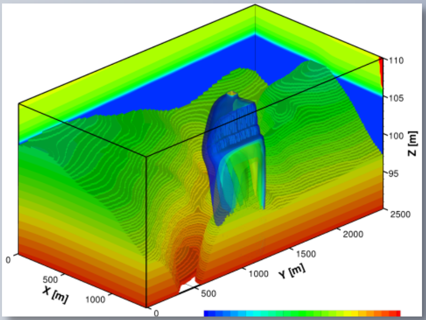


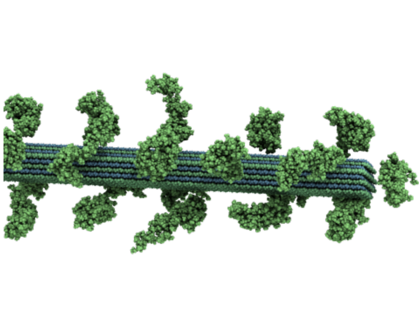


S3D
How are going to efficiently burn next generation diesel/bio fuels?

.

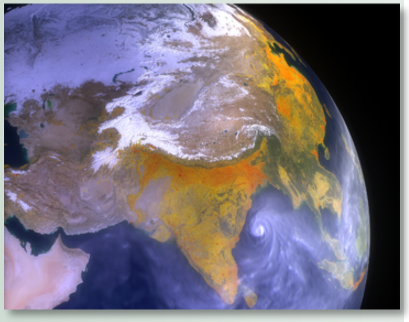
PFLOTRAN
Stability and viability of large scale CO_2 sequestration; predictive containment groundwater transport

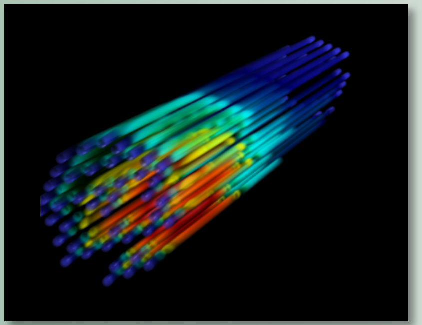




LAMMPS
Simulated time evolution of the atmospheric CO_2 concentration originating from the land’s surface

CAM / HOMME
Answer questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns/statistics and tropical storms





Denovo
Unprecedented high -fidelity radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

Titan Project: Programmer Productivity

- **Code team for each project**
  - ☑ Science team, performance engineer, applied mathematician, library specialist
- **Working with vendors on tools**
  - ☑ CAPS (HMPP) – Compiler mods for accelerators, C++
  - ☑ Allinea – Scale DDT to 250K cores; support for accelerators
  - ☑ Vampir – Support for profiling accelerator code
  - ☑ Cray – Compilers, Performance tools, unified tool set
- **Application Readiness Review of our preparation**
  - ☑ Spent 6 months analyzing and porting 6 applications to a hybrid CPU/GPU platform
  - ☑ Review panel was asked to assess our analysis of the challenges, level of effort, and potential for performance gains of science applications on a hybrid architecture

Results of Application Readiness Review of Titan Accelerator-Based system

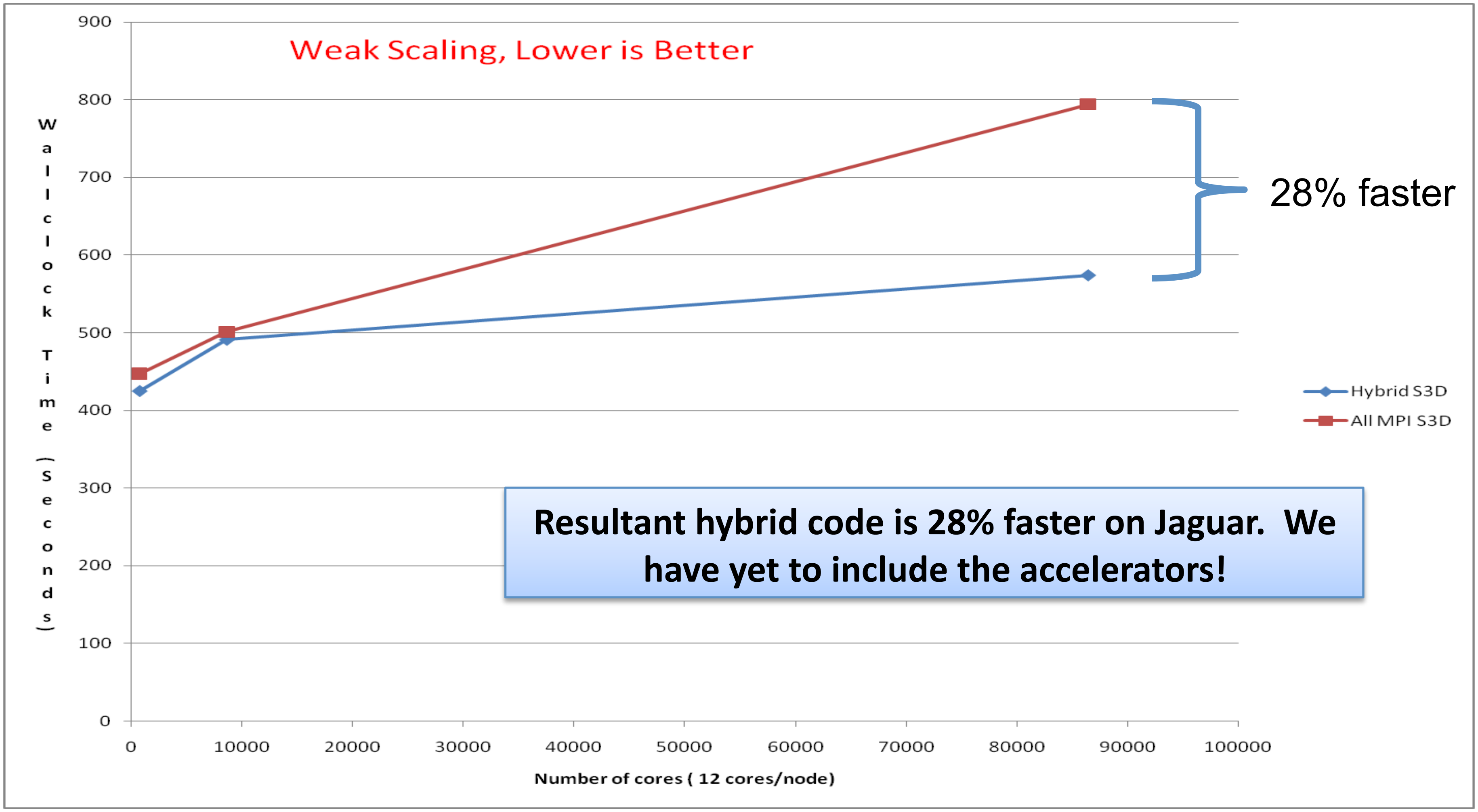
- “Use of the GPU did lead to a performance relative to power cost improvement in almost all cases.”
- “There is significant upside potential in GPU performance as we learn how to effectively use manycore architectures and develop new algorithms.”
- “GPUs are a harbinger of all future processors to come and there is ample evidence that designing applications for today’s GPUs will positively impact the performance of all multicore and manycore processors both today and in the future.”
- “Giving OLCF users access to a machine that is competitive as both a CPU and GPU system will provide an excellent transition vehicle for manycore applications development.”

Case Study: Hybridization of S3D

- **Objective: Restructure an all MPI application to utilize shared memory parallelism and vectorization on the node**
- **Reasoning: With such an application structure, the resultant code can efficiently run on existing multi-core systems as well as future hybrid systems**
- **Process: Identify areas in the program where high level loops can be introduced to give high granularity parallel structures**
    - Introduce grid loops within time step loop
    - Separate message passing from computation
    - Overlap communication with computation
- **Result: Hybrid MPI/OpenMP application developed that has**
    - Better Cache Utilization
    - Better Vectorization at low level
    - Reduction of required Memory
- **Next Step: Employ OpenMP accelerator extensions to put high level OpenMP structures on the accelerator for OLCF3**

Slide courtesy of John Levesque

Resultant Hybrid S3D Performance



Resultant hybrid code is 28% faster on Jaguar. We have yet to include the accelerators!

Slide courtesy of John Levesque

Early Science Applications on OLCF-3

- The six apps described above will be only part of the initial early science vanguard on Titan
- A RFP will be used by the OLCF sometime later this CY
- Early Science Apps will be selected based on a variety of factors
  - Scientific impact
  - Alignment with DOE SC missions
  - Current application readiness
  - Maturity of hybridization development plan
- More details to come soon
- Selected teams should expect time on the IDS at some point

Questions?

Bronson Messer

Email: bronson@ornl.gov

The research and activities described in this presentation were performed using the resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.